CS584 – machine learning

fall 2016

Regression Model for

Predicting Movie Box Office Gross

Group Members:

Ying Wu (A20370189)

Yingjuan Wu (A20326320)

Sahand Zeinali (A20318383)

Table of Contents

[Task 2](#_Toc467172373)

[Dataset 2](#_Toc467172374)

[Data source 2](#_Toc467172375)

[Target variable 2](#_Toc467172376)

[Features 2](#_Toc467172377)

[Data size 2](#_Toc467172378)

[Preprocessing 2](#_Toc467172379)

[Visualization 2](#_Toc467172380)

[Target 2](#_Toc467172381)

[Features 2](#_Toc467172382)

[Evaluation 2](#_Toc467172383)

[Performance Measure 2](#_Toc467172384)

[Classifiers 3](#_Toc467172385)

[Evaluation Strategy 3](#_Toc467172386)

[Performance Results 3](#_Toc467172387)

[Top Features 3](#_Toc467172388)

[Discussion 3](#_Toc467172389)

[Interesting/Unexpected Results 3](#_Toc467172390)

[Contributions of Each Group Member 3](#_Toc467172391)

[Conclusion 3](#_Toc467172392)

[References 3](#_Toc467172393)

**Regression Model for Predicting Movie Box Office Gross**

Group Members: Ying Wu, Yingjuan Wu, Sahand Zeinali

# Task

Is there a way to determine the greatness of a movie before it is released? How do we decide if a movie is worth watching instead of relying on instincts or critics? To answer this question, we explored a Kaggle dataset [1] which scrapes more than 5000 movies from IMDB website. Each movie has 28 features including its box office gross. In this project, we are trying to build a regression model which could predict a movie’s box office gross using relevant features available before a movie is released.

# Dataset

This dataset is available online [1]. It contains 28 features for 5043 movies spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors or actresses. Figure 1 shows the data for the first 20 movies in this dataset.

|  |
| --- |
|  |

Fig. 1 Screen shot of the first 20 movies in raw dataset

## Data source

We used this raw dataset on Kaggle directly, which scraped from IMDB website. We didn’t collect additional data or manually label any data since this dataset is large enough with a lot of information to be processed.

## Target variable

Our target variable is the box office gross of movies. We intend to predict this value based on regression method.

## Features

Apart from gross, which we selected as target variable, there are other 27 features in the datasets. These features are: movie\_title, color, num\_critic\_for\_reviews, movie\_facebook\_likes, duration, director\_name, director\_facebook\_likes, actor\_3\_name, actor\_3\_facebook\_likes, actor\_2\_name, actor\_2\_facebook\_likes, actor\_1\_name, actor\_1\_facebook\_likes, genres, num\_voted\_users, cast\_total\_facebook\_likes, facenumber\_in\_poster, plot\_keywords, movie\_imdb\_link, num\_user\_for\_reviews, language, country, content\_rating, budget, title\_year, mdb\_score, aspect\_ratio.

## Data size

There are 5403 movies and each movie has 28 features.

# Preprocessing

To preprocess the dataset, we did the following five things:

1. Remove redundant data. All duplicated movies are deleted.
2. Remove irrelevant features.

Features collected after a movie is released are removed, like aspect\_ratio and num\_user\_for\_reviews.

Features that are hard to be interpreted are removed, including movie\_title, movie\_imdb\_link, plot\_keywords, color, director\_name and actor\_name .

Features that are repetitive others are removed. For example, actor\_facebook\_likes is removed because we have cast\_total\_facebook\_likes, language is removed because country represents the same thing with more classifications.

1. Remove movies which were released before 1960 since they are too old and their features may be inaccurate.
2. Deal with missing values.

For movies that have missing gross or director\_facebook\_likes, they are removed.

For movies that have missing num\_critic\_for\_reviews, duration, facenumber\_in\_poster or budget, their missing value is replaced with the mean of all other movies.

For movies that have missing content-rating, their content-rating is set to “Not rated”.

1. Binarize categorical feature: Genres, Country and Content-rating.

After preprocessing, we have a clean dataset with 3321 rows and 87 columns.

# Visualization

## Target

<Provide statistics about the target variable; if classification, provide counts for each class; if regression, print mean and variance.>

<Provide a histogram of the target variable.>

## Features

<Visualize the features. For categorical variables, provide bar plots, for numerical features, provide histograms and statistics. If there are too many features, visualize and provide statistics for only 10 of them that you choose. If your data is text, print the list of most frequent words. If your data is raw (like images), describe how your model is handling it.>

# Evaluation

## Performance Measure

<Which performance measure you chose and why?>

## Classifiers

<Which classifiers and parameter settings did you try and why?>

## Evaluation Strategy

<Did you do train-test split or cross-validation and why?>

## Performance Results

<Report your results, including baselines, using a table similar to the one on slide 7 of the project presentation template file.>

## Top Features

<Present the top features with respect to your model.>

## Discussion

<Briefly discuss your results. Did the best classifier perform as well as you expected? If things did not work out as well, why do you think they did not work? Did one classifier perform much better (or worse) than others? And so on.>

# Interesting/Unexpected Results

<Discuss a few interesting/unexpected cases. See slide 9 of the project presentation template file.>

# Contributions of Each Group Member

|  |
| --- |
| Ying Wu:  Evaluation:  Load processed data, Print data shape |
| Performance measures description |
| Report performance of baselines |
| Build linear regression model using different parameter settings and report MSE, R2 |
|  |
|  |
| Yingjuan Wu:  Data Exploration:  Binarize all categorical features, i.e. Genres, Country and Content-rating  Visualize five numerical features  Print Mean and Variance of all numerical features  Evaluation: |
| Continuous features scaling |
| Build ridge regression model with different parameters and report MSE, R2 |
| Build Bayesian ridge regression model with different parameters and report MSE, R2 |
| Report top features and their weights  Presentation and Report:  Add two slides in Presentation. Complete the first three sections, i.e. Task, Dataset, Preprocessing, in the report. |
|  |
|  |
| Sahand Zeinali:  Data Exploration:  Visualize target variable  Visualize five features (numerical and non numerical)  Visulize using bar plots and histograms  Evaluation: |
| Regression model use research |
| Build Lasso regression model with different parameters and report MSE, R2 |

Presentation and Report:

Added slides into the presentation regarding the formatting.

Visualization and Evaluation sections of the reports were added.

# Conclusion

<Provide concluding remarks.>

# References

[1] Kaggle dataset. Available [online] <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

Notes (erase these notes before you save and submit):

1. Once you are done editing, this file, update Table of Contents; you can simply click on Table of Contents and click on the button Update Table once it appears. Then, save this file as a pdf file.
2. Do not copy text from any website/paper. All the wording has to be your own.
3. Delete instructions (the text that appear inside <> above) and these notes before you save and submit.
4. Anywhere it says “discuss”, one or two paragraphs should be sufficient. Do not write anything unnecessary, but also do not omit any necessary details.